УДК: 57.02.001.57

ИСПОЛЬЗОВАНИЕ КЛАСТЕРНОГО АНАЛИЗА И ЛОГИСТИЧЕСКОЙ РЕГРЕССИИ ДЛЯ ДИФФЕРЕНЦИАЦИИ ПАТОЛОГИЙ ЛЕГКИХ

Д. Ю. Козлов

Алтайский государственный университет, г. Барнаул

Статья поступила в редакцию 19 ноября 2016 г.

Аннотация. В данной работе рассмотрена возможность использования кластерного анализа и логистической регрессии для дифференциации патологий (рак и туберкулез), приводящих к возникновению шаровидных образований в легких. Диагностическими признаками выбраны параметры, определенные на основе обработки изображений рентгеновской компьютерной томографии, а именно: среднее значение денситометрического показателя Хаунсфилда и среднеквадратичное отклонение денситометрического показателя Хаунсфилда для выделенной врачом-рентгенологом области интереса, а также фрактальная размерность и величина «уклон». Для полученной выборки последовательно были применялись кластеризация методом k-means и иерархическая кластеризация методом полной связи. При сопоставлении результатов кластерного анализа с верифицированными диагнозами был сделан вывод, что иерархическая кластеризация более надежно, чем метод k-means выделяет верный диагноз. Затем, на основе половины исходной выборки получена модель логистической регрессии. С помощью ROC-анализа оценивалось качество полученной модели, а также определялась пороговая точка отсечения, позволяющая перейти от вероятностей диагноза, полученных после применения логистической регрессии, к прогнозу самих диагнозов. После этого выбранная модель логистической регрессии использовалась для предсказания диагнозов во второй половине выборки, причем корректность прогноза оказалась около 70%.

Ключевые слова: шаровидные образования в лёгких, фрактальная размерность, кластерный анализ, логистическая регрессия.

Abstract. In this paper we have considered the possibility of using cluster analysis and logistic regression to differentiate pathologies (cancer and tuberculosis), leading to the appearance of spherical formations in the lungs. The parameters determined on the basis of imaging X-ray computed tomography (the average value of the densitometry Hounsfield index and standard deviation of Hounsfield index for the area of interest was selected by physician radiologist, and also the value of the fractal dimension and the "slope") was chosen as diagnostic features. To this sample, in series, clustering by k-means and complete linkage method for the hierarchical clustering were applied. When comparing the results of cluster analysis with verified diagnoses, it was concluded that the hierarchical clustering is more reliable than the k-means method for the correct diagnosis elects. Then, on the basis of half of the original sample a logistic regression model was obtained. ROC-analysis was used to estimate the quality of the resulting model, and also for the determination of cut-off point that allows to pass on from probabilities of diagnosis obtained after applying logistic regression to forecast own diagnoses. The selected logistic regression model was used to predict the diagnosis in the second half of the sample, and level of correct prediction about 70% was reached.

Keywords: spherical formation in the lungs, fractal dimension, cluster analysis, logistic regression.

Поспроцессинговая диагностика по медицинским изображениям зачастую очень сложна, она требует от врача высокой квалификации и большого опыта. В работах [1-7] исследовались медицинские изображения, полученные с помощью рентгеновской компьютерной томографии. У обследованных были обнаружены шаровидные образования в легких, которые могут возникать при следующих заболеваниях: рак, инфильтративный туберкулез и пневмония. И даже для опытного врача-рентгенолога представляет большую трудность различить по томографическому изображению две патологии: рак и туберкулез. В качестве помощника врачу-

диагносту могла бы выступать интеллектуальная информационная система, работающая на основе объективных числовых параметров. Соответственно, необходимо подобрать такие числовые характеристики, которые бы помогли различить эти два заболевания.

Рентгеновские компьютерные томографы могут сохранять результаты обследования в формате DICOM-файла. Программное обеспечение, использующееся при поспроцессинговом исследовании полученной томограммы, предоставляет врачу-диагносту среднее значение денситометрического показателя Хаунсфилда и среднеквадратичное отклонение денситометрического показателя Хаунсфилда для выбранной области.

Рис. 1. Изображение легких, полученное с помощью рентгеновской компьютерной томографии. Область интереса дополнительно выделена квадратом.

Для проведения исследования нам требовалась база файлов с известными диагнозами. Были отобраны 2490 изображений для пациентов с уже ранее верифицированными диагнозами, среди которых оказалось 1850 случаев рака и 640 случаев инфильтративного туберкулеза. Дополнительно, на томографическом изображении врач-рентгенолог выделял «область интереса» (см. рис. 1), которая сохранялась в отдельный файл со значениями денситометрического показателя Хаунсфилда каждого пиксела этой области.

Кроме повсеместно применяемых описательных статистик (среднего значения H и среднеквадратичного отклонения σ денситометрического показателя Хаунсфилда для области интереса), дополнительно вычислялись величины, являющиеся результатами фрактального анализа изображения, при котором в качестве меры была выбрана площадь. Затем в дважды логарифмическом масштабе строилась зависимость логарифма меры $M(\varepsilon)$ от логарифма масштаба ε этой области [7-8]. В итоге, производная $B = \frac{d\ln(M(\varepsilon))}{d\ln(\varepsilon)}$, названная «уклон», и фрактальная размерность D области интереса [8] дополнили список параметров для последующего анализа.

На этом этапе важно понять, правда ли выбранные величины по отдельности, либо совместно позволяют различить исследуемые патологии. В работе [7] было получено, что для рака и туберкулеза средние значения выборок показателей Хаунсфилда H и фрактальной размерности D статистически различны, поэтому их можно использовать для диагностики. А средние значения совокупностей среднеквадратичного отклонения денситометрического показателя Хаунсфилда σ и функции «уклон» B не могут считаться существенными диагностическими признаками, т.к. они статистически не различимы. Мы же попытаемся провести кластерный анализ и логистическую регрессию для дифференциации рака и туберкулеза. Входными параметрами будут среднее значение показателя Хаунсфилда H и среднеквадратичное отклонение денситометрического показателя Хаунсфилда σ, а также величина «уклон» B и фрактальная размерность D для отдельного

изображения. Быть может, эти четыре параметра совместно представляют большую диагностическую ценность, чем только две величины показателя Хаунсфилда *H* и фрактальной размерности *D*, как указано в [7].

Для разделения объектов на группы по близости характеризующих их параметров в самых разных отраслях [9-13] широко применяется кластерный анализ [9]. В нашем случае необходимо было выделить две группы по числу патологий (рак и туберкулез), поэтому процедура кластерного анализа немного упрощалась.

Сначала был применен метод кластеризации k-means (k-средних), алгоритм которого работает на основе минимизации суммарного квадратичного отклонения точек кластеров от центроидов этих кластеров [9]. Так как выбор исходных центроидов кластеров происходит случайным образом, метод k-means может сходиться не глобальному, а локальному минимуму суммарного квадратичного отклонения, из-за чего могут возникать неодинаковые кластеры от одного запуска метода к другому [14]. Поэтому, описанные ниже итоги кластеризации методом k-means являются результатом усреднения по большому числу запусков.

Кластеризация проводилась с использованием следующих наборов параметров: (*H*, *D*), (*H*, *D*, *B*), (*H*, *D*, σ) и (*H*, *D*, *B*, σ). Однако итог кластерного анализа методом k-means оказался не слишком удовлетворительным:

1. для набора (*H*, *D*), т.е. только среднее значение денситометрического показателя Хаунсфилда *H* и фрактальная размерность области интереса *D*, из 2490 наблюдений лишь в 1032 случаях (или в 41 проценте) результат кластеризации совпал с верифицированным диагнозом.
2. для набора (*H*, *D*, *B*), т.е. только среднее значение денситометрического показателя Хаунсфилда *H*, фрактальная размерность области интереса *D* и величина *B* («уклон»), получился аналогичный предыдущему результат – из 2490 наблюдений лишь в 1032 случаях (или в 41 проценте) результат кластеризации совпал с верифицированным диагнозом.

3. для набора (*H*, *D*, σ), т.е. фрактальная размерность области интереса *D*, среднее (*H*) и среднеквадратичное (σ) значения денситометрического показателя Хаунсфилда, – из 2490 наблюдений лишь в 1456 случаях (или в 58 процентах) результат кластеризации совпал с верифицированным диагнозом.
4. для полного набора параметров (*H*, *D*, *B*, σ), т.е. фрактальная размерность области интереса *D*, «уклон» *B*, среднее (*H*) и среднеквадратичное (σ) значения денситометрического показателя Хаунсфилда, получился аналогичный предыдущему результат – из 2490 наблюдений лишь в 1456 случаях (или в 58 процентах) результат кластеризации совпал с верифицированным диагнозом.

Часто при кластерном анализе применяется иерархическая кластеризация [9, 11, 12]. При такой кластеризации точки объединяются на основе матрицы подобия (в этой работе она строилась с помощью метода полной связи или метода дальнего соседа). В результате получается граф без циклов, называемый дендрограммой. При большом объеме входных данных дендрограмма не отличается наглядностью и поэтому здесь не приводится. После процедуры иерархического кластерного анализа получено следующее:

1. для набора (*H*, *D*), т.е. только среднее значение денситометрического показателя Хаунсфилда *H* и фрактальная размерность области интереса *D*, из 2490 наблюдений лишь в 1075 случаях (или в 43 процентах) результат кластеризации совпал с верифицированным диагнозом.
2. для набора (*H*, *D*, *B*), т.е. только среднее значение денситометрического показателя Хаунсфилда *H*, фрактальная размерность области интереса *D* и величина *B* («уклон») – из 2490 наблюдений в 1357случаях (или в 54 процентах) результат кластеризации совпал с верифицированным диагнозом.
3. для набора (*H*, *D*, σ), т.е. фрактальная размерность области интереса *D*, среднее (*H*) и среднеквадратичное (σ) значения денситометрического показателя Хаунсфилда, – из 2490 наблюдений в 1823 случаях (или в 73

процентах) результат кластеризации совпал с верифицированным диагнозом.

4. для полного набора параметров (*H*, *D*, *B*, σ), т.е. фрактальная размерность области интереса *D*, «уклон» *B*, среднее (*H*) и среднеквадратичное (σ) значения денситометрического показателя Хаунсфилда, – из 2490 наблюдений в 1618 случаях (или в 65 процентах) результат кластеризации совпал с верифицированным диагнозом.

Из приведенных результатов ясно, что методы иерархической кластеризации дают более высокий процент совпадений с верифицированным диагнозом. Также становится понятно, что не стоит окончательно отбрасывать возможность использования величин *B* и σ для диагностики.

Попробуем проверить это заключение другим способом. Разделим имеющуюся выборку на две равные части, содержащие по половине диагнозов каждого вида. Затем на основе первой из созданных таким образом выборок попробуем подобрать оптимальную модель логистической регрессии, которую далее используем для предсказания диагноза во второй половине исходной выборки.

Логистическая регрессия – один из распространенных методов бинарной классификации [9-12, 15, 16], который удобно использовать и в нашем случае для определения типа нозологии. Результатом логистической регрессии является вероятность принятия зависимой переменной одного из двух возможных значений (в нашем случае, рак или туберкулез) при определенном наборе значений независимых переменных. Перейти от полученных вероятностей к прогнозу конкретного диагноза, можно выбрав точку отсечения, т.е. такое пороговое значение вероятности, которое и позволит выделить искомое заболевание. Установить такой порог помогает ROC-анализ [10, 12]. Он изначально создавался для обработки радиолокационных сигналов (ROC расшифровывается как Receiver Operating Characteristic или рабочая характеристика приёмника), но затем нашел применение и в других областях.

Далее в этой работе строились модели логистической регрессии, в которых зависимая переменная определялась следующей совокупностью независимых переменных (предикторов): фрактальная размерность области интереса *D*, «уклон» *B*, среднее (*H*) и среднеквадратичное (σ) значения денситометрического показателя Хаунсфилда. Рассматривались варианты без взаимодействия предикторов между собой и при наличии такого взаимодействия. Затем при помощи дисперсионного анализа (ANOVA) и критерия хи-квадрат [15] сравнивались модели без взаимодействия и с взаимодействием независимых переменных. В результате, было выяснено, что все предикторы (фрактальная размерность области интереса *D*, «уклон» *B*, среднее (*H*) и среднеквадратичное (σ) значения денситометрического показателя Хаунсфилда) значимы, и лучше соответствует данным вторая модель с взаимодействием предикторов.

Следующим шагом стало использование полученной на основе первой половины исходной выборки модели логистической регрессии для предсказания вероятностей диагноза для второй половины исходной выборки. Потом была построена ROC-кривая, приведенная на рис. 2.

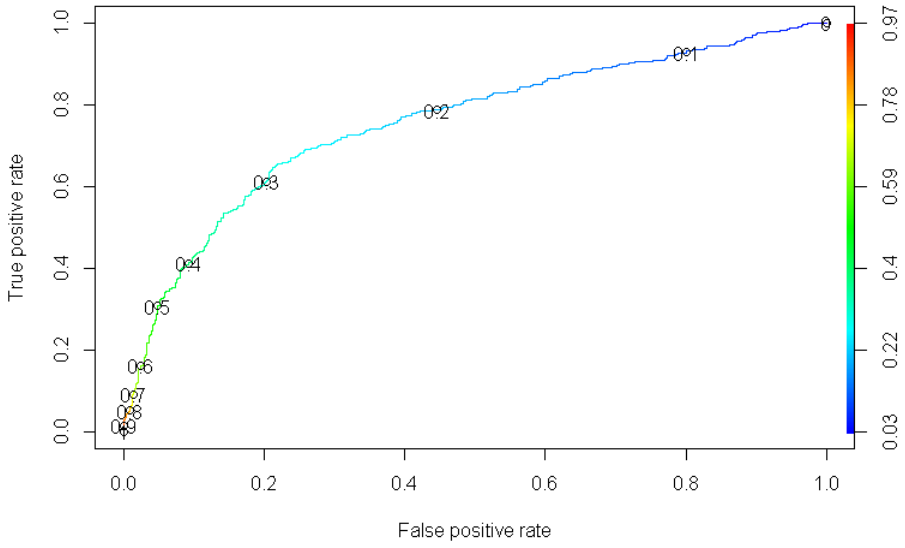

Рис. 2. ROC-кривая.

ROC-кривая в нашем случае показывает зависимость количества верно классифицированных диагнозов одного заболевания (True positive rate) от количества неверно классифицированных диагнозов другого заболевания (False positive rate). Площадь под кривой – параметр AUC = 0.76 – говорит о предсказательной силе модели, причем AUC = 1 соответствует идеальному классификатору, а значение AUC из интервала 0.7-0.8 считают обеспечивающим достаточно высокую точность [10].

Затем были построены кривые зависимости от порога отсечения следующих величин: чувствительности (в нашем случае – доля правильно определенных диагнозов одного заболевания), специфичности (в нашем случае – доля правильно определенных диагнозов другого заболевания) и точности (показатель эффективности классификации с помощью выбранной модели). В качестве точки отсечения было выбрано значение 0.25, соответствующее пересечению всех трех кривых (Рис. 3).

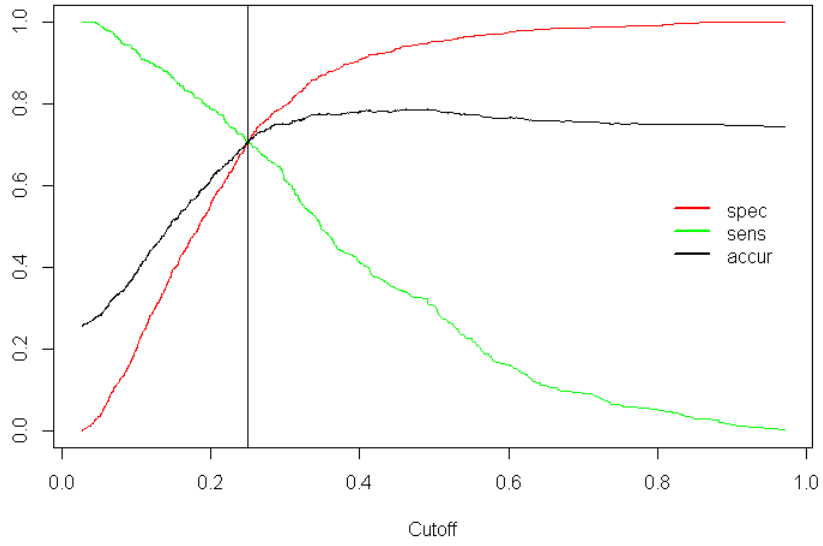

Рис. 3. Кривые чувствительности (sens), специфичности(spec) и точности (accur), а также вертикальная линия, соответствующая значению 0.25.

Теперь можно использовать точку отсечения для перехода от вероятностей диагноза к прогнозу самого диагноза. Были получены следующие результаты: предсказанный моделью логистической регрессии диагноз в 879 из

1245 наблюдений или в 71 проценте случаев совпал с верифицированным диагнозом.

Выводы

Из проведенного исследования понятно, что величины фрактальная размерность области интереса, функция «уклон», среднее и среднеквадратичное значения денситометрического показателя Хаунсфилда для области интереса можно применять для разделения трудно дифференцируемых патологий, приводящих к образованию шаровидных образований в легких. С точки зрения совпадения с верифицированным диагнозом величины «уклон» и среднеквадратичное значение денситометрического показателя Хаунсфилда улучшают корректность работы кластерного анализа и логистической регрессии. Для выделения верного диагноза иерархическая кластеризация подходит гораздо лучше метода k-means. Логистическая регрессия и иерархическая кластеризация почти одинаково классифицируют заболевания (71% и 65% соответственно) по полному набору предикторов.

Литература

1. Леонов С.Л., Шойхет Я.Н., Коновалов В.К. и др. Анализ погрешностей данных при мультиспиральной компьютерной томографии шаровидных образований легких // Проблемы клинической медицины. — 2011. — № 3-4 (25). С. 16-19.
2. Коновалов В. К., Шойхет Я. Н., Федоров В. В. и др. Прицельная 3d-реконструкция при изучении качественных характеристик поверхности шаровидных образований легких // Проблемы клинической медицины. — 2011. — № 3–4 (25). С. 20-25.
3. Шайдук А. М., Останин С. А., Коновалов В. К. и др. Проблема стандартизации масштаба при вычислении фрактальной размерности медицинских изображений // Известия Алтайского государственного университета. 2012. № 1-1 (73). С. 233-235.

4. Коновалов В.К., Шойхет Я.Н., Федоров В.В. и др. Способ прицельной объемной денситометрии шаровидных образований легких для оценки их внутренней структуры при мультиспиральной компьютерной томографии // Проблемы клинической медицины. 2012. № 1-4 (26-29). С. 74-86.
5. Коновалов В.К., Шойхет Я.Н., Федоров В.В. и др. Метод количественной оценки структуры шаровидных образований легких при мультиспиральной компьютерной томографии // Проблемы клинической медицины. 2012. № 1-4 (26-29). С. 95-101.
6. Останин С.А., Шайдук А.М., Козлов Д.Ю. и др. Энтропийный метод оценки сложности контура медицинских изображений // Известия Алтайского государственного университета. 2013. № 1-2 (77). С. 177-180.
7. Молодкин И.В., Леонов С.Л., Шайдук А.М. и др. Статистический анализ влияния типа патологии на количественные характеристики медицинских изображений // Медицинская физика. 2014. № 3 (63). С. 43-47.
8. Oczeretko E., Borowska M., Kitlas A. et al. Fractal analysis of medical images in the irregular region of interest.//BioInformatics and BioEngineering, BIBE 2008. 8th IEEE International Conference on Dept. of Med. Inf., Univ. of Bialystok, Bialystok, 2008. October.
9. Барсегян А.А., Куприянов М.С., Холод И.И. и др. Анализ данных и процессов: учеб. пособие — 3-е изд., перераб. и доп. — СПб.: БХВ-Петербург, 2009. — 512 с.
10. Паклин Н.Б., Орешков В.И. Бизнес-аналитика от данных к знаниям – СПб.: Питер, 2013. – 704 с.
11. Шипунов А.Б., Балдин Е.М., Волкова П.А., Коробейников А.И., Назарова С.А., Петров С.В., Суфиянов В.Г. Наглядная статистика. Используем R! – М.: ДМК Пресс, 2012. – 298 с.
12. Многомерный статистический анализ в экономических задачах: компьютерное моделирование в SPSS: Учебное пособие / Под ред. И.В. Орловой. - М.: Вузовский учебник, 2009. – 309 с.

13. Лесовых С.В., Тужикова Н.В., Юдинцев А.Ю. и др. Методика определения интегрального показателя уровня регионального развития // Тенденции науки и образования в современном мире. 2016. № 16-1. С. 39-43.
14. Бериков В.С., Лбов Г.С. Современные тенденции в кластерном анализе // Всероссийский конкурсный отбор обзорно-аналитических статей по приоритетному направлению «Информационно-телекоммуникационные системы», 2008. — 26 с.
15. Кабаков Р.И. R в действии. Анализ и визуализация данных в программе R / пер. с англ. П.А. Волковой. – М.: ДМК Пресс, 2014. – 588 с.
16. Мастицкий С.Э., Шитиков В.К. (2014) Статистический анализ и визуализация данных с помощью R. – Электронная книга, адрес доступа: http://r-analytics.blogspot.com